IN THE CLAIMS:

1. (Previously presented) A method for reducing magnitudes of output traffic bursts in a streaming media cache, comprising:

receiving a request from a first client system for a stream of media data, the stream of media data including a first streaming media data packet and a second streaming media data packet;

receiving a request from a second client system for the stream of media data;

receiving the first streaming media data packet from an upstream server, the first streaming media data packet including a delivery time;

pseudo-randomly selecting a first delay value and adding the first delay value to the delivery time of the first streaming media data packet to form a first modified delivery time for the first streaming media data packet;

pseudo-randomly selecting a second delay value and adding the second delay value to the delivery time of the first streaming media data packet to form a second modified delivery time for the first streaming media data packet;

modifying the first streaming media data packet with the first modified delivery time to form a first modified first streaming media data packet;

modifying the first streaming media data packet with the second modified delivery time to form a second modified first streaming media data packet;

outputting the first modified first streaming media data packet to the first client system at the first modified delivery time; and

outputting the second modified first streaming media data packet to the second client system at the second modified delivery time.

2. (Previously presented) The method of claim 1 wherein pseudo-randomly selecting the first delay value comprises pseudo-randomly selecting the first delay value from within a specified time range.

3. (Previously presented) The method of claim 2 wherein the time range is 0 to approximately 500 milliseconds.

4. (Canceled)

5. (Original) The method of claim 1 further comprising storing a payload portion of the first streaming media in a storage within the streaming media cache.

6. (Previously presented) The method of claim 2 wherein the second streaming media data packet includes a delivery time, the method further comprising:

adding the first delay value to the delivery time of the second streaming media data packet to form a first modified delivery time for the second streaming media data packet;

adding the second delay value to the delivery time of the second streaming media data packet to form a second modified delivery time for the second streaming media data packet;

modifying the second streaming media data packet with the first modified delivery time to form a first modified second streaming media data packet;

modifying the second streaming media data packet with the second modified delivery time to form a second modified second streaming media data packet;

outputting the first modified second streaming media data packet to the first client system at the first modified delivery time; and

outputting the second modified second streaming media data packet to the second client system at the second modified delivery time.

7. (Canceled)

8. (Original) The method of claim 1 further comprising:

receiving a data file from the upstream server, the data file including a payload portion of the first streaming media data packet and a payload portion of the second streaming media data packet; and

storing the data file in a storage within the streaming media cache.

9. (Previously presented) A computer system for providing streaming media data to client systems with reduced magnitude traffic bursts, comprising:

a first thread configured to receive a request from a first client system and a second client system for a stream of data packets, wherein the stream includes a first data packet and a second data packet, the first thread also configured to pseudo-randomly select a first client delay and to pseudo-randomly select a second client delay;

a second thread coupled to the first thread, the second thread configured to receive the first data packet and the second data packet from an upstream server, wherein the first data packet specifies a first delivery time and the second data packet specifies a second delivery time, the second thread also configured to form a first delayed first data packet from the first data packet based on the first client delay and to form a second delayed first data packet from the first data packet based on the second client delay, wherein the first delayed first data packet specifies a first delayed delivery time and the second delayed first data packet specifies a second delayed delivery time;

a third thread configured to receive the first delayed first data packet and to provide the first delayed first data packet to the first client system in response to the first delayed delivery time; and

a fourth thread configured to receive the second delayed first data packet and to provide the second delayed first data packet to the second client system in response to the second delayed delivery time.

10. (Original) The computer system of claim 9 wherein the second thread is configured to form the first delayed first data packet in response to the first client delay by adding the first client delay to the first delivery time.

11. (Previously presented) The computer system of claim 10 wherein the first client delay is pseudo-randomly selected from the range: 0 to approximately 500 milliseconds.

12. (Original) The computer system of claim 9 further comprising a thread configured to store payload portions of the first data packet and payload portions of the second data packet in a memory.

13. (Previously presented) The computer system of claim 9 further

wherein the second thread is also configured to form a first delayed second data packet in response to the first client delay and to form a second delayed second data packet in response to the second client delay, wherein the first delayed second data packet specifies a first delayed delivery time and the second delayed second data packet specifies a second delayed delivery time;

wherein the third thread is also configured to receive the first delayed second data packet and to provide the first delayed second data packet to the first client system in response to the first delayed delivery time specified therein; and

wherein the fourth thread is also configured to receive the second delayed second data packet and to provide the second delayed second data packet to the second client system in response to the second delayed delivery time specified therein.

14. (Original) The computer system of claim 13 wherein the second thread is also configured to form the first delayed second data packet in response to the first client delay by adding the second client delay to the second delivery time.

15. (Canceled)

16. (Previously presented) A method for reducing peak output traffic bursts in a processing system where a first packet of data is scheduled to be delivered to more than one downstream client system substantially simultaneously, the method comprising:

modifying a specified packet delivery time of the first packet of data for delivery of the first packet of data to a first downstream client system, by pseudo-randomly selecting a first delay value and adding the first delay value to the specified packet delivery time of the first packet of data; and

modifying the specified packet delivery time of the first packet of data for delivery of the first packet of data to a second downstream client system, by pseudo-randomly selecting a second delay value and adding the second delay value to the specified packet delivery time of second first packet of data.

17. (Original) The method of claim of claim 16 wherein the first packet of data is framed.

18. (Original) The method of claim 16 wherein the first packet of data comprises streaming media data.

19. (Previously presented) The method of claim 16 wherein pseudo-randomly selecting the first delay value comprises pseudo-randomly selecting the first delay value from within a specified time range.

20. (Previously presented) The method of claim 19 further comprising modifying a specified packet delivery time of a second packet of data for delivery of the second

packet of data to the first downstream client system, by adding the first delay value to a specified packet delivery time of the second packet of data.

21. (Previously presented) The method of claim 19 wherein pseudo-randomly selecting the second delay value comprises pseudo-randomly selecting the second delay value from within the specified time range.